Ceusters W. Formal terminology management for language-based knowledge systems: resistance is futile. In Temmerman R. (ed) Trends in Special Language and Language Technology, 2001;:135-53

Formal terminology management for language based knowledge systems: resistance is futile.

Dr. W. Ceusters

Abstract: The traditional - i.e. non-formal - approach to terminology management, including multi-lingual approaches supported by computers, is focused around the appropriate use of terms in the right context, and the internal organisation of terminology collections by means of relations such as narrower than, broader than, part-of, etc. Such an organisation makes a terminology manageable for humans and has proved to be useful for better translations, easier and more accurate writing of documentation, and so on. However, this approach is insufficient when demands are more complex such as automated verification of very large terminology collections or semantic-based document processing. This is where formal terminologies come into play: they are primarily designed to be understood by machines, and not by humans.

In this paper, we address specifically the need for formal terminologies in extremely large domains such as healthcare. The number of concepts in healthcare is estimated to be in the range of 10 to 20 million, a collection that never can be maintained in the traditional way. But when properly managed by using the right tools, a wealth of possibilities becomes available to overcome the burden of today's knowledge overload.

1. Language based knowledge systems

In business and manufacturing, three components have been considered extremely important: people, money and resources. Recently, a fourth component has been added: knowledge. In the consulting business, it has even become the most important component of all.

What is knowledge, and where does it come from? A traditional view is the knowledge production cycle. At the beginning, there are raw unprocessed data. Once collected and formatted, they can be processed in such a way that relationships become visible: data are turned into information that can be used to improve business or manufacturing processes. The more information that is disclosed, the more clever we become, until so much information (inside a domain) is disclosed, that a level of deep understanding is reached: the knowledge level. Having this knowledge, it becomes easier to derive more and better information from the data. We might even control the events that produce the data. Becoming knowledgeable is itself a matter of moving from one state into another (the “knowledge microprocess”): observation, understanding, prediction, application, justification. The ultimate goal (hopefully) is to reach the level of wisdom. This level can only be reached when sufficient knowledge from various domains has been acquired.

Out of this insight, various technologies have emerged. Information Technology can be defined as a supporting technology to turn data into information. Knowledge Engineering is a similar discipline at the level of information and knowledge. Language Engineering can be seen as a special branch of knowledge engineering, dealing with knowledge in the form of language. Finally, Language Based Knowledge Systems are systems in which most of the knowledge is stored in free text as in Document Management Systems, but that differ from the latter in their ability to actively - perhaps even “consciously” - participate in the knowledge production cycle. As such, Language Based Knowledge Systems have a sufficient level of "understanding" of the domain covered by the texts they contain, to assist users in adequately storing and retrieving information.

2. Healthcare as a knowledge intensive environment …

1 … with respect to size

Medicine is one of these complex domains where new knowledge is accumulated at a daily basis, and at an exponential rate. Most of this knowledge resides in textbooks and papers, or more loosely structured in patient records. Despite the growing tendency to make this knowledge available in electronic format, the era of large scale knowledge based systems does not seem to have dawned yet. On the one hand, we are perhaps close, issues such as complexity and expressive power of knowledge representations being better understood. On the other hand, there is still a very long way to go as representing large quantities of knowledge is a major bottleneck if we ever want to develop systems that don't "fall of the knowledge cliff" [1].

One of the research domains that might come up with solutions for the knowledge acquisition bottleneck is natural language understanding. A necessary condition is however that systems could be build that transform sentences into a meaning representation that is independent of the subtleties of linguistic structure that nevertheless underlie the way language works [2]. Yet, here also is a bottleneck, be it in the form of a chicken and egg problem. Systems showing this much wanted behaviour must base their inferences on knowledge already available to them. The knowledge required comes in two different flavours. First, there is that kind of knowledge that is described as "linguistic semantics", the rules and principles explaining how literal meaning is grammaticalised or encoded in language [3]. It is this kind of knowledge that enables us for instance to understand the "sense" of an expression or sentence, i.e. the set (or network) of sense-relations that hold between it and other expressions within the same language [4], and that allows us to identify the same meaning independent of whether a given sentence is in the passive or active form. This knowledge is different from "conceptual semantics" or "conceptual knowledge" that describes what entities there are (supposed to be) in the world that can be denoted by language.

2 … with respect to communication needs

The pharmaceutical industry is well aware of the importance of effective knowledge and information management. Bringing a new drug to the market is a multi-stage process that typically takes between 7 and 15 years. Huge amounts of information have to be gathered, analysed and communicated. Between 100 and 1000 people intervene somewhere in the drug development life cycle: feasibility studies, planning, clinical trial monitoring, medical writing, regulatory affairs, post-marketing surveillance, pharmacovigilance, etc. Tens of thousands of documents are generated and have to be analysed.

The pharmaceutical industry is also a multi-national business. This means not only that multiple languages have to be addressed, but also that effective communication channels have to be set up across the language borders.

Communicating information is an essential activity in the whole healthcare domain. Patients need to be informed by their doctors to what diagnostic or therapeutic procedures they will be submitted in order to make them feel more comfortable in the often threatening environment of the hospital. Also nurses need to be informed on what happened to the patients there are responsible for before taking up their shift. It is mandatory that this exchange of information is done in an ambiguous, accurate and reproducible way. This is not always so easy because language itself - the prime vehiculum in information interchange - is difficult to use unambiguously. In addition, Europe is moving towards a global multilingual community in which from a functional perspective, national borders tend to fade. More often communication is required with colleagues speaking different languages, or having another cultural and educational background.

Given the rather limited capacities of the human brain in storing and retrieving large quantities of factual data, the same information must also be registered in patient records for subsequent consultation. By using electronic patient records, some additional functional requirements for this kind of “external memories” became apparent: in one way or another, the information has to be understandable by machines, such that linking to other applications or information sources can be achieved nearly automatically. Unfortunately, computers don’t speak natural language (yet), and they also have little knowledge of medicine.

To overcome the problems related to the use of natural language in medical communication and clinical registration, terminology collections in the form of coding and classification systems have been introduced as interlingua. Systems such as ICD, Snomed International, ICPC, CPT and many others are now widely used to register medical findings, diagnoses or procedures. Similarly, so called terminological systems such as NIC, NANDA, ICNP and others are proposed to be used as interlingua in a nursing environment.

The question of course is whether or not such systems are the right solutions to overcome the problems stated previously. After all, each of these systems is designed with a specific purpose in mind such as mortality and morbidity statistics, reimbursement, information retrieval to mention only three. They very seldom are detailed enough for a faithful registration of all relevant clinical data. And at least in their current (paper) format, they are a burden to use.

3. Non-formal terminology

In [5], terminology is defined as the study and the field of activity concerned with the collection, description, processing and presentation of terms belonging to specialised areas of usage of one or more languages. Central in this definition is the notion of terms, i.e. verbal representations of the things we speak or write about. Terminology differs from lexicology in the sense that only the terms pertaining to a specific domain are considered.

Three dimensions need to be considered when developing terminologies: the cognitive dimension, the linguistic dimension, and the communicative dimension. Also when existing terminologies are to be compared to be used in a specific environment, it is mandatory to keep these dimensions in mind.

In the cognitive dimension, the terms are related to their conceptual contents, i.e. the referents in the real world independent of their material or abstract nature. In this dimension, terms get their meaning fixed. In the linguistic dimension, the existing and potential forms of the terms are examined. Here term formation principles are studied.

The communicative dimension finally looks at the use of terminologies. This dimension has to justify terminology work as such.

A rigorous method must be adopted when designing terminologies. Also, it is mandatory that the work is undertaken by a multidisciplinary team composed of skilled terminologists, linguists, and domain specialists. Usually, one starts by defining the area of usage, the application domain and the intended purpose. If a multilingual terminology is aimed for, also the source- and target languages need to be identified. As a first step, large corpora of documents need to be collated. These documents might be other terminologies developed within the domain under scrutiny - perhaps for a different purpose - or texts in which a high number of candidate terms can be found. This approach is justified when it is assumed that if a term is found in a document (the linguistic dimension), there must be a concept that it denotes (the cognitive dimension). On the basis of this material, a taxonomy of the terms can be set up, i.e. identifying generic relationships between them. Studying the taxonomy might in itself give clues for the existence of concepts for which no terms have been found in the initial corpus.

Special care needs to be taken when doing the work in a multilingual environment. It is always dangerous to translate terms directly from one language into another without giving careful thoughts at the concepts they denote. When the meaning of a term in language A is not exactly equal to the meaning of another term in language B, both terms should not be considered to be each other’s translation.

4. Towards formal terminologies

Non-formal terminologies (nomenclatures, thesauri, classifications, etc.) are designed to be used by humans. Even electronic versions of these systems, in which it is possible to browse through the hierarchies of the terminology, are still intended to be used by humans, the computer just being there as a replacement for the book. A major problem for such naïve electronic versions is that they cannot take advantage of the knowledge implicitly available in the terms (or the rubrics in classification systems), but that they must rely on the limited knowledge available in the generic links between terms. Finding specific terms requires a priori knowledge by the user on how the system is structured. With flat terminologies, in which large quantities of narrower-terms depend from one broader-term, the computer is even seen as a burden, because only a limited number of terms can be seen at the same time on the screen. A second disadvantage is that the terminologies only can be viewed in their original structure, and that reclassification of the terms, following different criteria, cannot be realised.

In order to overcome these problems, terminologies must be expressed in a formal way. When doing so, the three dimensions of terminology should not be forgotten.

1 Formalisation along the cognitive dimension

The cognitive dimension takes care of the meaning of terms. Traditionally, meanings in particular domains are found in specialised dictionaries, i.e. large books meant to be used by humans to look up the meaning of unknown words. Most electronic dictionaries currently available differ only from paper dictionaries in that they are published on a digital medium. To be useful in Language Based Knowledge Systems, dictionaries have to be fundamentally different in nature: they are primarily meant to be used by machines!

The important thing in these dictionaries, is that entries are not related to each other directly, but through a language independent formal concept system. An important characteristic of such a system is the clear separation of different kinds of relationships that hold between the concepts. “Language independence” does not mean ignoring language as a medium of communication, a mistake quite often committed by people working in that field, but being independent from any particular language [6]. If the concept system is solely intended to be used as a knowledge base for internal processing, without any communication being needed in natural language, then there are some arguments for such an approach. If not, it will definitely lead to unsatisfactory behaviour. The good approach is to keep the concept system separate from any linguistic knowledge. But in addition, a linguistic ontology is to be maintained, capturing the relationships between the grammars of particular languages, and the language independent concept system. This is explained further down.

For humans, it is sufficient to define *Zenker’s diverticulum* as a *diverticulum of the oesophagus caused by intraluminal pressure*, to make the term meaningful. An electronic dictionary intended to help human readers, may be implemented as a 2-column table, the first containing the terms, the second containing the definitions.

For a machine, this format is totally unacceptable. Definitions need to be dissected completely, while each building block must have a meaning on its own. Meaning is added to dictionary entries through explicit context definition. A number of knowledge building blocks must be defined, of which “concepts” and “linktypes” are the most important ones. Concepts refer to things that may be instantiated in the real world, while links relate concepts amongst each other.

A clear distinction should be maintained between IS-links as formal subsumption relations and other links. This guarantees that automatic classification of newly defined concepts can be achieved, freeing the knowledge engineer from the need to give manually new concepts the most accurate place(s) in the concept system.

Once such a component is available in a Language Based Knowledge System, it is possible to develop various applications. Simple ones are keyword or word-spotting based and can be used for automated encoding, an activity that is extremely important in Europe where in many countries, the revenues of hospitals depend on coding medical diagnoses and procedures. Other applications allow medical staff to highlight sentences or paragraphs in patient documents such as discharge letters to access bibliographic services, even in other languages.

2 Formalisation along the linguistic dimension

.1 There is language and language ...

Formalising terminologies along the conceptual dimension is “all” that is needed to allow computers to make properly use of them. It is however not sufficient if communication is required between computers and humans, and certainly not for interpersonal communication. The former requires a mapping from the formal language to a language understandable by humans and vice-versa, while the latter requires the unambiguous use of natural language amongst humans.

While formal language and natural language are at the two most extremes of an axis representing the understandability of a language for a machine or a human respectively, there are two kinds of languages that more or less can bridge the gap.

The first kind encompass "sublanguages", i.e. natural languages used in a particular domain, f.i. nursing, and for a particular task, f.i. communicating or documenting nursing interventions. The second one are known as "controlled languages". A controlled language is a precisely defined subset of a natural language, on the one hand constrained in its lexicon, grammar and style, and on the other hand possibly extended by domain-specific terminology and grammatical constructions. Both controlled languages and sublanguages have in common that they differ from "general" natural languages by being restrictive, deviant and preferential with respect to vocabulary, syntax, semantics and pragmatics [7, 8, 9, 10]. The main difference is however that sublanguages evolve naturally within a community while controlled languages are artificial adaptations of a language that are tried to be kept as natural as possible. Controlled languages are not to be mixed up with "controlled vocabularies" that are (possibly hierarchically) structured sets of certified terms that are verbal canonical representations of concepts. The aspect of control in a controlled vocabulary is related to the position of a specific term in the vocabulary as a whole, the choice of a particular term as canonical form, and the requirement that only terms from within the vocabulary are to be used in an application. The terms themselves are however not written in a controlled language. In [11], we proposed the use of a controlled language to reduce ambiguity in the terms or rubrics of medical nomenclatures, vocabularies, and coding and classification systems (Table 1). This was based on the many inconsistencies and ambiguities that were found in Snomed International [12] (Table 2).

Table 1: Some basic recommendations for controlled language usage in term formation for clinical nomenclatures

1. Avoid using the same word in different meanings and with different parts of speech.
2. Use prepositions in such a way that they (preferably uniquely) identify the thematic role or object-relation.
3. Use double or triple prepositions for expressing meaning with greater precision.
4. Maintain normal word order as indicated by the general grammar of the language in which the terms are expressed.
5. Limit term length to what (at least) a skilled human reader can easily understand.
6. Use co-ordination with extreme care.

Table 2: Phenomena reducing the understandability of terms in Snomed International

1. Inappropriate use of synonymy
2. Misleading use of homonyms
3. Complexity of noun groups or noun clusters
4. Long-distance dependency and cross-modification of term constituents
5. Ambiguous use of co-ordinated constructions
6. Different (and unpredictable) semantics of the word "and".

.2 Cognitive versus linguistic modelling

When formalising terminologies along the cognitive dimension, an *ontology* has to be defined, i.e. a representation - to be used in computer systems - of what concepts exist in the world, and how they relate to one another. Ontologies are often viewed as strictly language independent models of the world, especially in the medical informatics community, though the need for an ontology in natural language processing applications is generally well accepted [13]. This is not to say that knowledge structuring based on a linguistic approach leads to the same result as when opting for a conceptual approach. A typical example is the ontological distinction between *nominal* and *natural kinds* [14], that in no language is grammaticalised just because the difference is pure definitional [15]. This again does not mean that such distinctions are not useful in a natural language processing applications.

Situated ontologies - i.e. ontologies that are developed for solving particular problems in knowledge based applications [16] - that have to operate in natural language processing applications, are better suited to assist language understanding when the concepts and relationships they are built upon, are linguistically motivated [17]. In the perspective of re-usability, two dimensions have however to be explored: (relative) independence from particular languages and (relative) independence from particular domains. Linguistic semantics based analyses allow us to separate f.i. entities from events and property concepts, a rather crude distinction being the fact that in most languages these concepts are respectively grammaticalised by means of nouns, verbs and adjectives [3]. Linguists are concerned on how these concepts give overt form to language, while from a computational point of view, these concepts also have to be “anchored” in a *linguistic ontology*.

While formalising medical terminologies along the conceptual dimension, numerous examples can be found where linguistic principles are in conflict with conceptual principles [18]. Physicians want to see medical concepts organised in a framework that reflects their clinical way of thinking. As an example, concepts such as “filling” and “injecting” can be categorised as specialisations of a “LiquidInstallingProcess” that itself is a child of “InstallingProcess”. This categorisation is useful from a clinical perspective where from the place in the hierarchy it can be derived that the concepts of injecting and filling have to do with the installation of liquid. This categorisation does however not line up with the linguistic structures that (at least in European languages) are used to express installing, filling and injecting events. From a language understanding perspective, it would be better to categorise these motion events according to the way the thematic roles of *goal* and *theme* may surface in sentences expressing these events. Also concerning part-whole relationships, there are differences in categorisation and actual expressions. Clinicians wants to have the fingernail classified as part of the upper extremity, following a long chain of transitivity over “distal phalanx”, “finger”, “hand”, “lower arm” and “arm”, while they would never actually say that “a fingernail is a part of the upper extremity”.

.3 Unifying the cognitive and linguistic dimension: the interface ontology approach

A relative new notion related to ontologies is that of the *interface ontology*, standing between conceptual (or domain) and linguistic ontologies. Approaches based on interface ontologies differ in the “distance” between the interface ontology and the domain ontologies at the one hand, and the linguistic ontologies at the other hand. In the MikroKosmos initiative, an interface ontology is developed for machine translation purposes in the domain of commercial merges and acquisitions of companies (19). Hence, it is more close to a given conceptual domain, although general concepts are included as well as unrestricted texts are envisaged to be processed. The KOMET project resulted in the "Generalised Upper Model 2.0", where a closer contact with linguistic realisations is maintained: *if there is no specifiable lexicogrammatical consequences for a 'concept', than it does not belong in the Generalised Upper Model* (20, p5). As a linguistically oriented ontology, the GUM is fundamentally different in design from domain- or world-knowledge oriented ontologies in that it captures those distinctions which have influences for grammatical expressions in distinct languages without committing to just what the grammatical distinctions of any particular language are. This therefore provides a powerful point of language localisation that maintains theoretical independence from particular linguistic theories and language engineering techniques.

A relatively similar, though more simple approach is used in EuroWordNet [21]. In this project, semantic databases like WordNet [22] for several languages are combined via a so-called inter-lingual-index (ILI). This allows language-independent data to be shared over the languages, while language-specific properties are maintained as well in each individual database. The only organisation provided to the ILI is via two separate ontologies. The first one is the top-concept ontology which is a hierarchy of language-independent concepts, reflecting explicit opposition relations. The second is a hierarchy of domain labels. Both the top-concepts and the domain labels can be transferred via the equivalence relations of the ILI to the language-specific meanings and, next, via the language-internal relations to any other meaning in the individual database of a specific language.

3 Formalising along the communicative dimension

It is often stated that concept systems in health care must be language- and purpose independent, and that they should be formally described in a powerful and expressive formalism on which computationally tractable algorithms can be applied. However, our analysis of the relevant literature in the domains of medical informatics, computational linguistics and philosophy has shown that these requirements cannot be fulfilled at the same time [23]. Language - independence cannot completely be achieved as structuring the knowledge domain and building the concept system is a matter of thematic sublanguage analysis and of subcategorisation which itself only can be performed by using the information provided in a given language. In different languages, the same concept may be subcategorised on different criteria or features.

Purpose - independence seems to be the most problematic goal to achieve as orientation towards a purpose is required for (1) identifying what concepts should be represented, (2) deciding on what should be introduced in the concept system as a concept or as a role, (3) eliminating unnecessary complexity of the concept system's structure by avoiding unneeded subcategorisations, and (4) limiting the depth of the terminology in order to avoid the problems associated with the computational intractable property of many formal terminological systems. The interest-relativity of conceptual systems is due to the fact that descriptions tend to have a particular explanatory role. When describing objects, answers to particular questions are implicitly given. What is accepted as an interesting answer, is usually a context-sensitive matter [24].

The communicative dimension of terminologies is both related with the maintenance of terminologies, and the purpose(s) for which they are designed. As a consequence, problems such as how to guarantee that a (formal) terminology is properly used for what it is designed for, how can it be put in practice, how can it be maintained, and what is needed to allow co-existence with other systems, need to be accounted for. To all these questions, there is one common answer: there must be a general computational framework upon which various terminological tools and applications can be built. Such a framework must be specifically designed for graph- and network operations such that it can be considered to be a database manager for knowledge represented in the form of a semantic network. API's can be developed to integrate the system in front-end applications. This computational framework is the kernel of any Language Based Knowledge System.

An important aspect of the communicative dimension of a specific terminology is its relationship with other terminologies in the same or a related domain, be it possibly developed for different purposes. Quite often, mapping tables are set up as a means to go from one terminology to another. Ideally however, all systems should be represented formally according to a common framework. This has the advantage that mapping tables are an automatic by-product of such an effort.

5. A case study: fifth generation electronic healthcare records as Language Based Knowledge Systems

Though most clinicians and other healthcare workers are gradually becoming convinced of the advantages of using computers, they still prefer to retrieve data stored by others, than to register data themselves. There are many reasons for this such as unavailability of systems at the point of care, incomplete integration in the primary care process, or the fact that only a subset of the activities for which clinicians would like to have computer support, are actually offered.

The issue that deserves our particular attention in this paper is the *information structuring bottleneck*. Healthcare records, whether on paper or in computers, are originally kept as an external record for individual patient histories, such that future decisions can be based appropriately on past events. Electronic patient record systems have additional advantages over paper-based systems in their ability to allow for cross-patient studies, and to provide active decision management functionalities. While the former requires thorough structuring of the data inside the machine, the latter also requires representing and storing knowledge and information in the machine so that the machine *itself* can manipulate it, at least for tasks for which it is better suited than humans.

The need for structured *data representation and storage* being undeniable and very well understood, the need for structured *data entry* seems to be the logical consequence. This is at least the impression that we get from analysing the data acquisition interfaces of so many electronic healthcare record systems. There is structuring at the level of the data capture modalities such as rigorous data entry forms, point and click interfaces, structured menu's, etc. There is also structuring at the level of content by using coding and classification systems or controlled vocabularies. The question should be whether or not it is necessary to require the structuring be done by the user. Or as Tange et al. phrase it: "*Initiatives to facilitate the entry of narrative data have focused on the control rather than the ease of data entry*" ([25], p. 24). It is a fact, that most users don't like structured data entry at all, but that many accept it in the light of the benefits obtained when retrieving information. They accept the burden of structured data entry as the price to be paid for powerful information retrieval. But is this price affordable, let alone justifiable ?

Many clinicians share the view that faithful recording of patient data can only be achieved by using natural language. This was already stated in the early eighties by Wiederhold who claimed that *the description of biological variability requires the flexibility of natural language and it is generally desirable not to interfere with the traditional manner of medical recording* [26]. Also more recently, strong arguments have been given to preserve natural language registrations in clinical records and to view them under a “narratological framework” as proposed by Kay and Purves [27].

Besides this theoretical and fundamental position in favour of natural language registration, there is also a practical reason: data entry by means of continuous speech recognition (CSR). CSR technology has now reached a functional threshold in transforming a speech signal into digital text what is all that is needed for dictation. However, inexperienced users quickly might infer from this evolution that all data entry could be done by voice, freeing them from the need to use a keyboard. Despite this demand, CSR is not that easy lined up with structured data entry forms or cascaded menu’s. The command and control paradigm for navigating through forms and menu’s is only acceptable in a “hands free” situation, but even that still requires visual feedback from the screen. The ideal situation would be one in which users can enter information or issue queries in natural language, upon which the machine would analyse and structure the input automatically. This calls for advanced natural language understanding.

Implementations of electronic patient record systems should find an adequate balance in dealing with clinical language, rigid database structures and medical terminologies. Unfortunately, this is not the case with the clinical data entry paradigms most systems adhere to today.

Many papers describe the kind of data that are to be registered in an EHCR, some from a standardisation perspective [28], others on more technical or scientific grounds [29]. Prior to define a framework for modelling the EHCR, a clinical account is given by Rector et al [30, 31]. An essential criteria is that the record should give a faithful account of the clinician’s understanding. Data should be formulated in terms that are found natural. Conflicting statements must be allowed and also uncertain and negative statements must be accepted. Descriptions should be given at any arbitrary level of detail and at the clinicians’ natural level of abstraction. Once entered, data should be there permanent. Though this description fits the characteristics of free text registration, the authors argue that also structured data entry paradigms should fulfil these requirements. Unfortunately, they never do.

A typical example is the ICPC (International Classification of Primary Care, currently being replaced by ICPC-2). It has proved to be a valuable tool for statistically comparing the activities of GP surgeries, based around the concept of “reason for encounter”. It consists of a small classification of around 780 terms that clearly cannot be used to describe all relevant information with respect to individual patient care. The same can be said of other coding and classification systems that try to generalise healthcare information by abstracting away from the details that are judged irrelevant for the specific purpose that they have been designed for. But irrelevant for a specific purpose, does not necessarily mean irrelevant for all individual patients.

The logical conclusion is: if systems are not designed for capturing all relevant data for individual patient care, then don’t use them for that purpose ! Hence, developers of electronic patient record systems that want to integrate these systems (usually only available as long lists without adequate searching facilities) into their applications have only one good option: the systems must be integrated in addition to other data entry facilities, and users must be instructed that it does not suffice to register a number of codes out of such systems to have a faithful recording. They'll have to register in free text, and then must assign codes afterwards for all systems that are required according to institutional or governmental directives. Only medical natural language understanding technology can improve this situation.

6. Conclusion

These are the facts that (in our view) dictate Language Based Knowledge System design in general, and electronic healthcare record systems in particular:

1) natural language is the only medium that is able to communicate (clinical) information about individual cases without loss of necessary detail;

2) structured data repositories are required to make subsequent analyses possible;

3) any transformation from free language to coding and classification systems results in information loss that is unacceptable. This is specifically the case for EHCR systems were information loss is unacceptable for individual patient care, but at the other hand is a conditio sine qua non for population based studies;

4) today's graphical user interfaces can deal reasonably well with picking lists build around controlled vocabularies that fulfil a bridging function from free language towards coding and classification systems. However, speech recognition technology will soon free the user from the screen, such that item selection isn't anymore an option.

5) User interfaces must be designed in such a way that they don't disturb the primary process. There must come a shift from the current paradigm of user-initiated "data-entering" towards machine-initiated "data-capture": the machine observes without any interference of what is going on.

To make this happen in the domain of healthcare, medico-linguistic ontologies will need to become essential components of any EHCR system. Medical ontologies that have been designed without keeping the language-constraints in mind, are doomed to fail: "*The current implementation of SNOMED-RT does not have the depth of semantics necessary to arrive at comparable data or to algorithmically map to classifications such as ICD-9-CM*" [32, p70]. The same goes for systems that are mainly build around language, without adequate conceptual design, such as is the case for UMLS and its components: "*Simply using everything in the Metathesaurus does not make a good coding system*" [33], and "*The problems with the Metathesaurus as a single monolithic vocabulary are: 1. There is a wide range of granularity of terms in different vocabularies, 2. The Metathesaurus itself has no unifying hierarchy, so you cannot take advantage of hierarchical relations, 3. There may be other features of vocabularies that get lost in their "homogenisation" upon being entered into the Metathesaurus.*" [34].

The only good approach is to have systems that keep natural language, structured representations and formal terminologies nicely in balance. That this is possible, has already been shown successfully [35, 36]. As such, a start has been made to improve on current implementations of computerised terminology database management systems, in line with, but independent from the principles of Sociocognitive Terminology [37].

7. References

1 R.S. Michalski, Understanding the nature of learning: issues and research directions, in: R.S. Michalski, J.G. Carbonell and T.M. Mitchell, eds., Machine Learning, an artificial intelligence approach, vol II (Morgan Kaufmann Publishers Inc., Los Altos, 1986) 3-25.

2 J. Allen, Natural Language Understanding (The Benjammin/Cummings Publishing Company Inc, Menlo Park California, 1987).

3 W Frawley, Linguistic Semantics (Lawrence Erlbaum Associates, Hillsdale, 1992).

4 J. Lyon, Linguistic Semantics, an introduction (Cambridge University Press, Cambridge, New-York, Melbourne, 1995).

5 Sager JC. A Practical Course in Terminology. John Benjamins Publishing Company, Amsterdam, 1990.

6 Ceusters W, Rogers J, Consorti F, Rossi-Mori A. Syntactic-semantic tagging as a mediator between linguistic representations and formal models: an exercise in linking SNOMED to GALEN. Artificial Intelligence in Medicine 1999; 15: 5-23.

7 Kittredge R., Lehrberger J. (eds.) : Sublanguage : studies of language in restricted domains. de Gruyter, Berlin, 1982.

8 Harris Z. Mathematical Structure of Language. John Wiley & Sons. New-York, 1968.

9 Harris Z. A theory of language and information. Clarendon Press, Oxford, 1991.

10 Ceusters W, Spyns P, De Moor G, Martin W (eds.) : *Syntactic-semantic tagging of medical texts: the MultiTALE-project.* IOS Press, Amsterdam, 1997.

11 Ceusters W, Steurs F, Zanstra P, Van der Haring E, Rogers J. *From a time standard for medical informatics to a controlled language for health*. International Journal of Medical Informatics 1998, 48: 85-101.

12 Côté R.A., Rothwell D. (eds.), *Systematized Nomenclature of Medicine - SNOMED International*, College of American Pathologists, Chicago, 1993

13 Bateman JA. Ontology construction and natural language. *In Proc. International Workshop on Formal Ontology.* Padua, Italy, 1993, 83-93.

14 Kripke S. Naming and Necessity. In Davidson D & Harman G (eds.) *Semantics of natural language*. Dordrecht: Reidel, 1972, 253-355.

15 Welsh C. *On the non-existence of natural kind terms as a linguistically relevant category*. Paper presented at the Linguistic Society of America, New Orleans, LA, 1988.

16 Mahesh K & Nirenburg S. A situated ontology for practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95.* Montreal, Canada, 1995.

17 Deville G, Ceusters W. A multi-dimensional view on natural language modelling in medicine: identifying key-features for successful applications. Supplementary paper in *Proceedings of the Third International Working Conference of IMIA WG6*, Geneva, 1994.

18 Ceusters W. *Language Engineering as an Enabling Technology for Clinical Terminology Harmonisation*. In: CEC-DGXIII (ed.) Important Issues in Today’s Telematics Research, TAP’98 Conference Barcelona, 1998, 168-173.

19 Mahesh K. 1996. *Ontology Development for Machine Translation: ideology and methodology*. Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

20 Bateman J, R. Henschel and F. Rinaldi. 1995. Generalised Upper Model 2.0: documentation, GMD/Institute for integrated publication and information systems Technical Report, Darmstadt, Germany.

21 Vossen P, P. Diez-Orzas, and W. Peters. 1997. The Multilingual Design of the EuroWordNet Database. in: Proceedings of the IJCAI-97 workshop on Multilingual Ontologies for NLP Applications, Nagoya, August 23.

22 Miller GA, R. Beckwidth, C. Fellbaum, D. Gross, and K.J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database, International Journal of Lexicography ¾, 235-244.

23 Ceusters W, Deville G, Buekens Ph. *The Chimera of Purpose- and Language Independent Concept Systems in Health Care*. In Barahona P, Veloso M, Bryant J (eds.) Proceedings of the XIIth International Congress of EFMI, 1994, 208-212.

24 Buekens F, Ceusters W, De Moor G. The Explanatory Role of Events in Causal and Temporal Reasoning in Medicine, *Met Inform Med* 1993, 32: 274 - 278.

25 Wiederhold G. Databases in healthcare. Stanford University, Computer Science Department, Report No. STAN-CS-80-790, 1980.

26 Tange HJ, Hasman A, de Vries Robbe PF, Schouten HC. Medical narratives in electronic medical records. International Journal of Medical Informatics, 1997, 46(1): 7-29.

27 Kay S, Purves IN. Medical Records and Other Stories: a narratological framework. Methods of Information in Medicine 1996; 35: 72-87.

28 Gabrielli ER. Standards for Electronic Patient Records. Journal of Clinical Computing, 20 (1), 1991, 21 - 32.

29 van Ginneken AM, tam H, Moorman PW. A multi-strategy approach for medical records of specialists. International Journal of Biomedical Computing 42: 1996, 21-26.

30 Rector AL, Nowlan WA, Kay S. Foundations for an electronic medical record. Meth Inform Med 30: 1991, 179-186.

31 Rector AL, Nowlan WA, Kay S, Goble CA, Howkins TJ. A framework for modelling the Electronic Medical Record. Meth Inform Med 32: 1993, 109-119.

32 Elkin PL, Harris M, Ogren PV, Buntrock ID, Brown SH, Solbrig HR, Chute CG: “Semantic Augmentation of Description Logic based Terminologies” Addendum to Proceedings of IMIA-WG6, Medical Concept and Language Representation, Phoenix, 16-19/12/1999, 70-81.

33 William T. Hole M.D., Director, Metathesaurus Research and Development, National Library of Medicine. Message to UMLS-users Mailing List, 23-06-2000.

34 Hersh W. Message to UMLS-users Mailing List, 23-06-2000.

35 Ceusters W, Laga M. *Introducing Language Engineering Tools to Support Information Processing in Healthcare Telematics*. In: Proceedings of Toward an Electronic Health Record Europe ’99, 14-17 November 1999, London (UK), 251-255, 1999.

36 Ceusters W, Lorré J, Harnie A, Van Den Bossche B. *Developing natural language understanding applications for healthcare: a case study on interpreting drug therapy information from discharge summaries*. Proceedings of IMIA-WG6, Medical Concept and Language Representation, Phoenix, 16-19/12/1999, 124-130.

37 Rita Temmerman. Towards New Ways of Terminology Description: the Sociocognitive approach.John Benjamins Publishing Company, 2000.